ITM Probe Documentation

Release 1.5.4

Aleksandar Stojmirovic

September 01, 2017

Contents

1 Overview

ITM Probe is a an application for modeling information flow in *protein interaction networks* based on discrete time *random walks*.

1.1 Information Flow Framework

Each random walk simulates a possible information path. A walker starts at a node in the network and at each discrete time step moves randomly to a node that is adjacent to its originating node. The probability of moving to a particular destination node is proportional to the weight of the link from the originating node to it. Unlike the classical random walks, our framework allows the walker a certain probability to *dissipate*, that is, to leave the network at each step. Each walk terminates either by dissipation or by reaching a boundary node.

We distinguish two types of boundary nodes: *sources* and *sinks*. Each source serves as origin of random walks and hence emits information, while each sink absorbs information. *ITM Probe* offers three models: *absorbing*, *emitting* and *channel*.

The **emitting model** contains only sources as boundary points. For each source s and any network node i, it returns the expected number of visits to i by a random walk originating at s. The **absorbing model** contains only sinks as boundary and for each network node j and each sink k returns the likelihood of a random walk starting at k to terminate at j. The **normalized channel model** contains both sources and sinks as boundary and combines the emitting and absorbing models. It reports the normalized expected number of visits to any node i from a random walk originating at a source s and terminating at sink k. The normalization excludes the walkers that do not reach any sink.

Each selection of boundary nodes and rates of dissipation provides the biological *context* for the information transmission modelled. A small dissipation coefficient allows random walks to explore the nodes farther away from their origin while a large one means that most walks will evaporate very quickly. Since the channel model considers only those walks that do not terminate due to dissipation, the large dissipation values will illuminate the nodes on the shortest paths from sources to sinks. The dissipation coefficients may be specified directly or through an *alternative criterion* (page 3).

We call the set of most significant nodes (with respect to the number of visits, in the case of the emitting and channel models, or the absorption probability, in the case of the absorbing model) an *Information Transduction Module* (*ITM*).

1.2 Alternative Ways of Specifying Dissipation

For the **absorbing model**, the alternative to specifying the dissipation coefficient directly is to set the average absorption probability. The average absorption probability is calculated by taking the probability of absorption at the sinks for every transient node connected to sinks, and averaging it by the number of transient nodes and the total number of sinks. It thus represents the likelihood of a random walk starting at a randomly selected point in the network to reach a sink. Every dissipation coefficient between 0 and 1 has a corresponding average absorption probability and hence specifying a valid absorption probability indirectly specifies the dissipation coefficient. The dissipation coefficient obtained in this way is larger if the sinks are well-connected hubs near the centre of the network as sinks, in contrast to the case when the chosen sinks are not as well connected.

In the case of the **emitting model**, the dissipation coefficient can be specified through the average path length. The average path length represents the average length of the path that a random walk originating at a source travels before dissipating. There is a one-to-one correspondence between path lengths and dissipation coefficients.

The node dissipation coefficient for the **channel model** can be specified through the amount of drift (deviation) from the shortest path length from sources to sinks. Since the channel model concerns only those random walks that originate at the specified sources and terminate at specified sinks, the minimum distance a random walk can take (at full dissipation) is the smallest length of the shortest path from one source to one sink. Allowing deviations from the shortest path corresponds to smaller dissipation. It is possible to express the drift from the shortest path in absolute or relative (in terms of the length of the shortest path) units.

1.3 ITM Probe Software

ITM Probe can be used through three interfaces: command line (*qmbpmn-tools* package), web and as a Cytoscape (http://www.cytoscape.org/) plugin.

The standalone version is written in the Python programming language and is a part of the *qmbpmn-tools* package, which contains additional utilities. The package has numerous dependencies including Numpy, Scipy, UMFPACK, Jinja2, Sphinx and Graphviz. The source code is made freely available but the package is in heavy development, lacks full documentation and is not supported in any way. Only the users who wish to examine the source code for *ITM Probe* algorithms or to reproduce the webpages of the QMBP molecular network tools (*ITM Probe* and *SaddleSum* locally) should download and attempt to install *qmbpmn-tools*.

The web version provides access to *ITM Probe* through a web form. The users are restricted to querying only the few compiled protein interaction networks from model organisms networks available on the website. The server-side scripts are part of the *qmbpmn-tools* package and rely the Graphviz suite for visualizing *ITM Probe* results.

Cytoscape is an open source platform for complex network analysis and visualization written in Java programming language. Apart for a rich set of graph visualization tools, it provides an interface for externally written plugins that provide additional functionality such as network analysis algorithms, database import and functional enrichment analysis. Cytoscape users are therefore able to combine algorithms and data from different sources to perform complex network-based analyses.

CytoITMprobe is a Cytoscape plugin that enables *ITM Probe* queries from Cytoscape platforms. It can interact either with a locally installed command-line program directly, or through the QMBP web server. Any Cytoscape network can be passed to *ITM Probe*. The output is written as the node attributes of the query network, which can be manipulated further within Cytoscape.

2 Web Service

The web interface for *ITM Probe* consists of a query submission page and a results page. The query page is divided into two panels: Model Parameters,and Display Parameters, while the results page contains Display Parameters and Results pannels, plus a form allowing submission of the results to *SaddleSum*, our term enrichment analysis tool. Each model (channel, emitting or absorbing) uses a different page for query submission, with a slightly different Model Parameters panel.

To run an *ITM Probe* query, first select the desired model page (from the main page). In the Model Parameters panel, select the interaction graph and excluded nodes (either by entering them directly or by uploading a text file with one node per line), and then set the contexts. When finished, press the **RUN** button to submit the query.

The output of *ITM Probe* consists of the graphical representation of the resulting ITM, query-related summary statistics and a table listing the top ranking members of the ITM. The number of members of ITM to be shown as well as the type of the graphical representation are controlled through Display Parameters panel, either on the query submission or on the results page. Thus, it is possible to run a single query and then explore different aspects of the results using different display options.

2.1 Model Options

All models require setting of the interaction network graph and the excluded nodes as well as the model-specific context.

Interaction Graph

While the mathematical framework of *ITM Probe* can be applied to any directed graph, the web service presently supports only the yeast (*Saccharomyces cerevisiae*), human (*Homo sapiens*), fruit fly (*Drosophila melanogaster*) and flatworm (*Caenorhabditis elegans*) physical interaction networks derived from the BioGRID (http://www.thebiogrid.org) database. We plan to offer more networks in future.

The *ITM Probe* web service offers three types of networks: **Full**, **Reduced** and **Directed**. Presently, all supported organisms have Full networks, while the Reduced and Directed types are available only for yeast as it has the network with by far the best quality and coverage. All physical interaction networks from the BioGRID contain the following interaction types: 'Affinity Capture-MS', 'Affinity Capture-RNA', 'Affinity Capture-Western', 'Co-crystal Structure', 'Co-fractionation', 'Co-localization', 'Co-purification', 'FRET', 'Far Western', 'Reconstituted Complex', 'Two-hybrid', 'PCA' and 'Biochemical activity'.

Full networks contain all interactions from the BioGRID for a particular species. All interaction types are treated as undirected and given a weight 1.0 independently of the number of publications reporting the same interaction.

Reduced networks are derived from Full by filtering out those interactions that are reported solely by high-throughput studies (that is, by publications reporting more than 300 interactions). The interactions reported by more than one publication are not filtered out. For yeast, we also do not filter the interactions reported by *Collins et al. (2007)* (PMID: 17200106) and by *Tarassov et al. (2008)* (PMID: 18467557), because they use newer experimental techniques (as opposed to two-hybrid) and because their reported interactions are

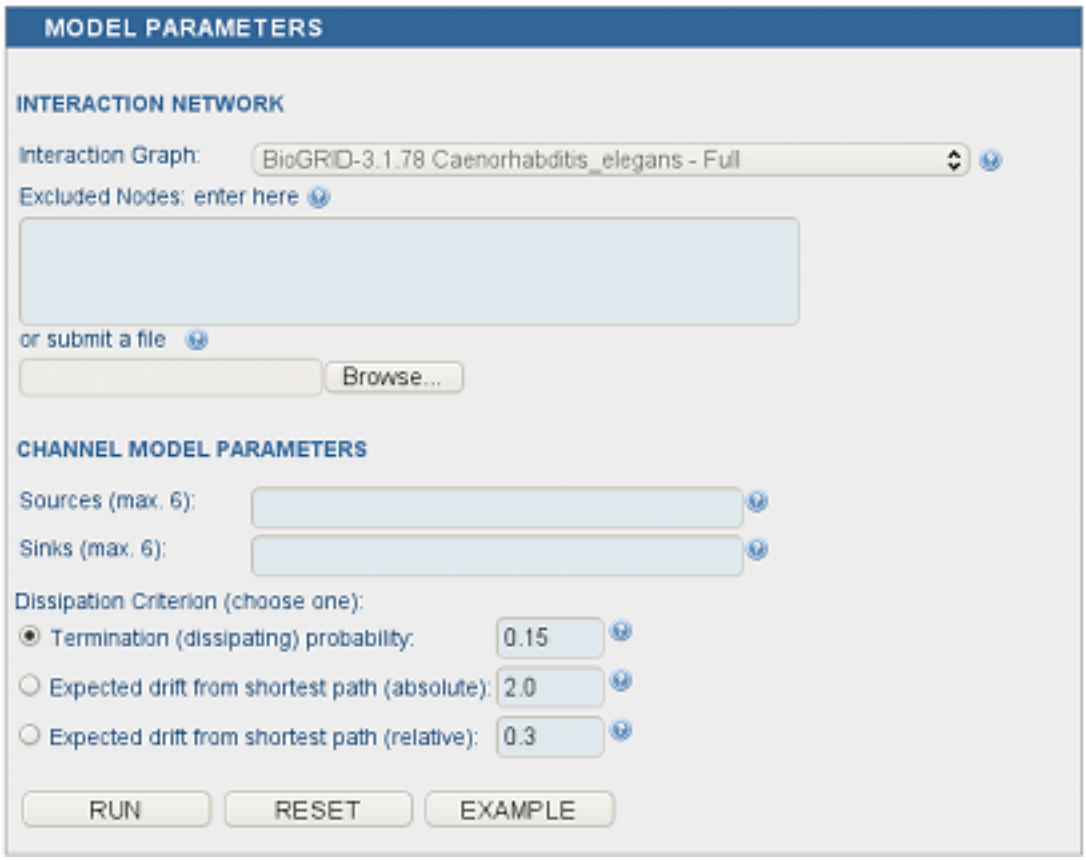
generally verified by more than a single experiment. As for Full networks, the edges of Reduced networks are all undirected with weight 1.0.

Directed networks are obtained from Reduced by turning all 'Biochemical activity' (phosporylation, ubiquitination etc.) interactions into directed links (bait to prey). Interactions are divided into two groups, 'Biochemical activity' ones and all others, and independently filtered by throughput as for Reduced network (hence we accept a 'Biochemical activity' interaction from a high-throughput study only if another 'Biochemical activity' link is reported by another publication for the same pair). Then, those interactions from the 'other' group that have an accepted corresponding 'Biochemical activity' link are removed. Finally the filtered 'Biochemical activity' interactions are introduced into the graph as directed links with weight 2.0 and all remaining interactions are taken as undirected with weight 1.0. In this way, Directed networks favor the 'Biochemical activity' and are therefore better than Full or Reduced for investigating potential signalling cascades.

The supported yeast networks also supply the list of about 170 *excluded nodes*, representing the cytoskeleton proteins, histones and chaperones that may provide undesirable shortcuts. The networks from other species do not have default excluded nodes but users can enter their own sets, if they wish so.

Channel Model

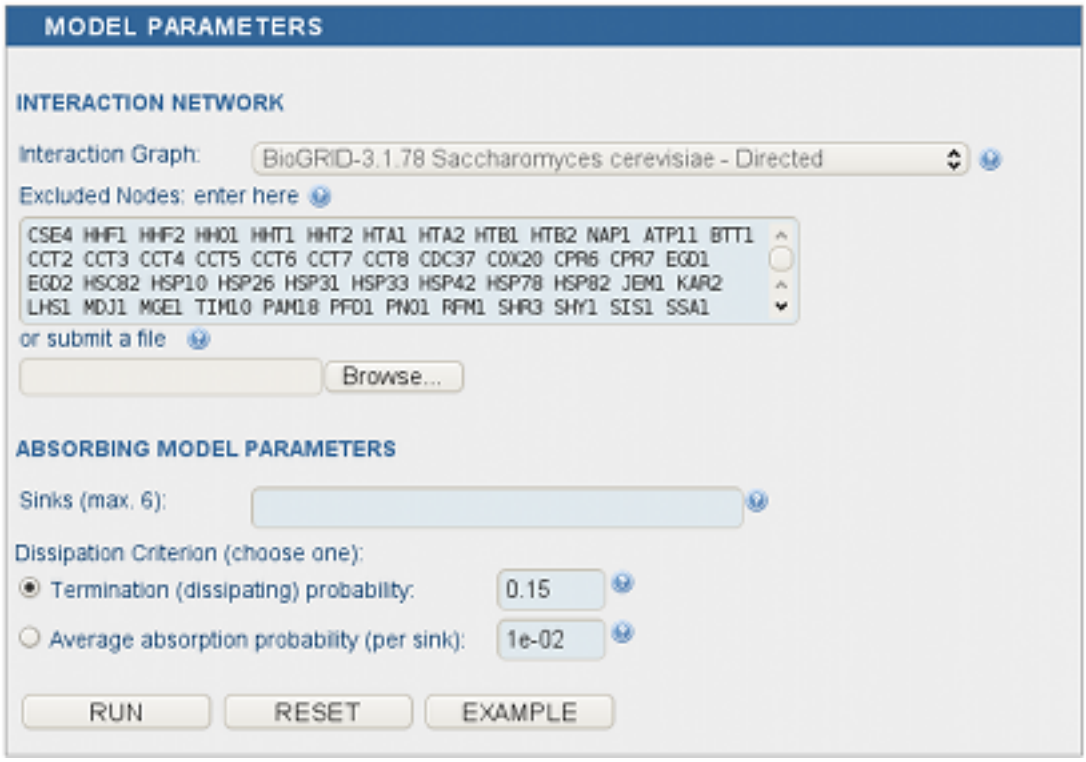
The (normalized) channel model requires a context, containing at least one source, one sink and a node dissipation coefficient between 0 and 1. It is possible to specify at most six sources and six sinks.



Channel model parameters panel.

Absorbing Model

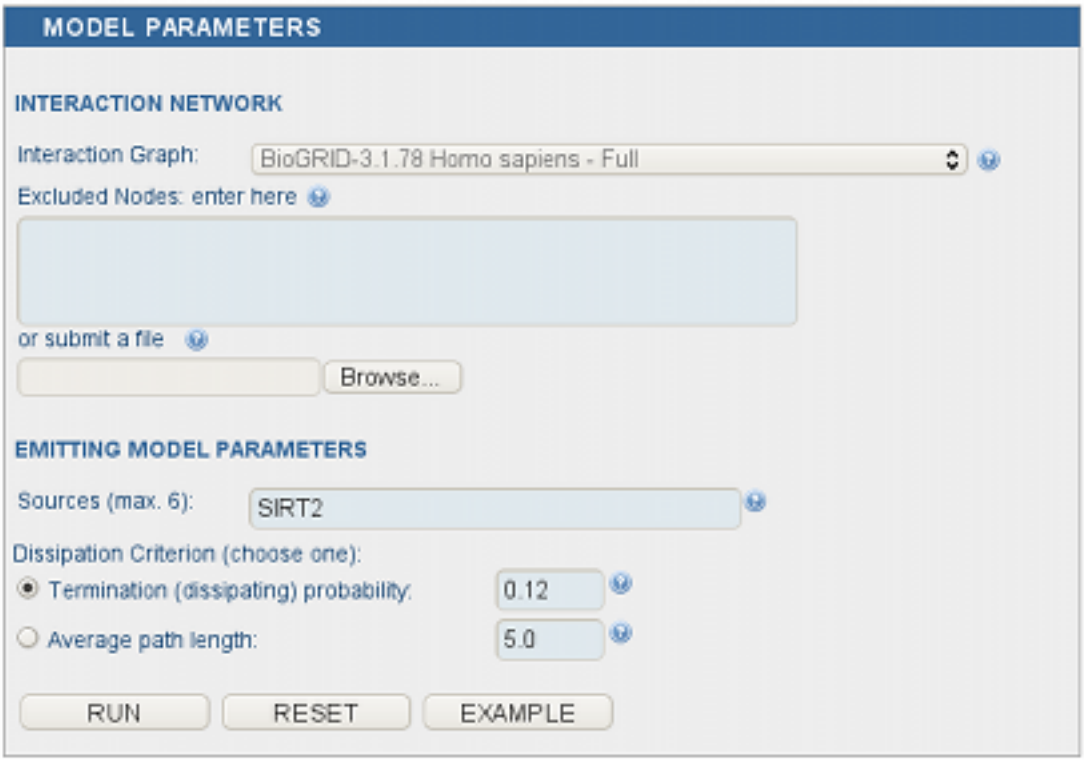
The absorbing model requires a context with one to six sinks and the dissipation coefficient.



Absorbing model parameters panel.

Emitting Model

The emitting model requires a context with one to six sources and a dissipation coefficient.



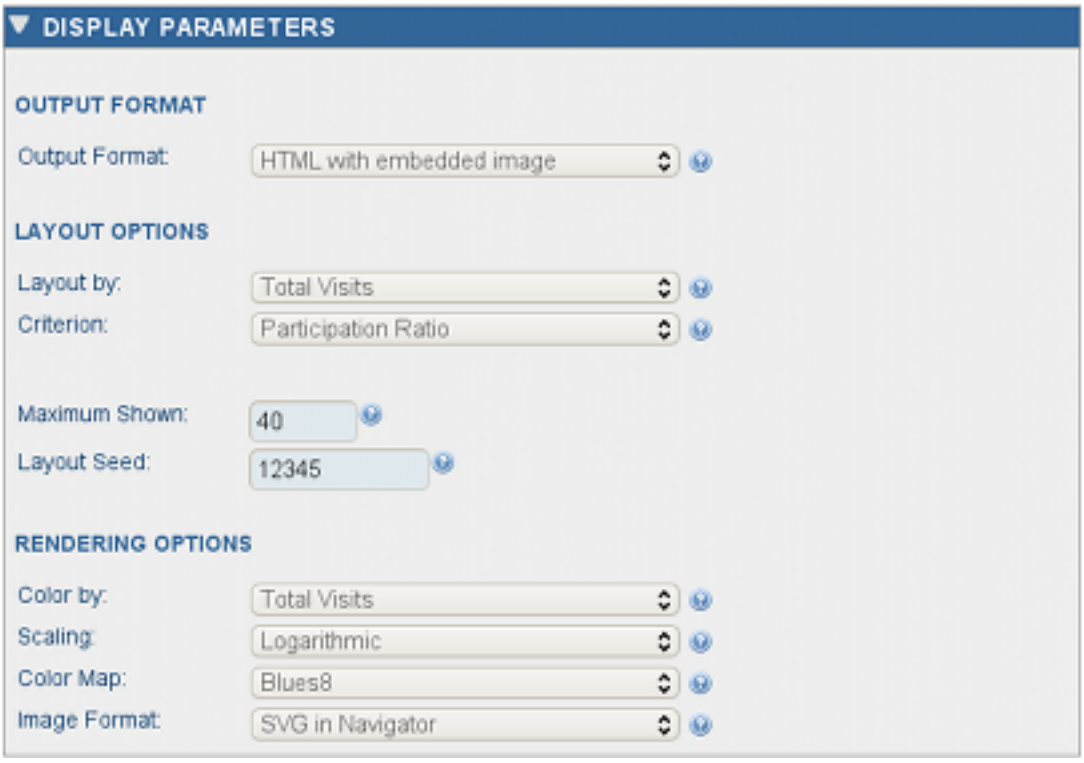
Emitting model parameters panel.

2.2 Display Parameters

Display parameters specify the way the evaluated context query is to be shown in the results panel. They include the output format, the layout options and the rendering options. The output format may be HTML, Plain Text or CSV. Only HTML format support embedded images of ITM subgraphs.

If an ITM image is present, the layout options control the number of most significant nodes shown and the layout of the nodes in image. The rendering options specify how a particular layout is colored and rendered. A single layout can be reused for different renderings, each showing different aspects of the ITM obtained from the query.

It is possible to specify the display parameters both on the query submission page and on the results page. To change the display on the results page, adjust the options in the Display Parameters Panel and hit the **RENDER** button. Only the image and the number of items in the table of results may change, the results of the query will remain the same.



Display parameters panel.

Layout options

The **Layout by** selection box specifies the choice of the ranking criterion for selecting the most significant nodes. At present, the choice is restricted to a *Total visits* and *Interference* for the channel and emitting models and *Total Likelihood* for the absorbing model.

The **Criterion** selection box specifies the criterion used to select the number of top ranking nodes to be shown. The default option for the channel and emitting models is *Participation ratio*. Other options are *Maximum nodes* (the fixed number of nodes specified by the **Maximum Nodes** option is shown) or *Cutoff Value* (the nodes having the associated value specified in the **Layout by** box greater than the **Cutoff Value** are displayed), the default for the absorbing model.

The **Maximum Nodes** option specifies the maximum number of nodes to be shown, capped by 200. This option overrides the number obtained from the **Criterion** selection in case of the *Participation ratio* or the number of nodes above the *Cutoff Value* are too large.

The placement of nodes in the graphical image is determined according to the **Layout Seed** option. Each seed produces a different layout and hence it may be possible to obtain a better image of and ITM by changing this parameter from its default value and trying several different seed values.

Rendering options

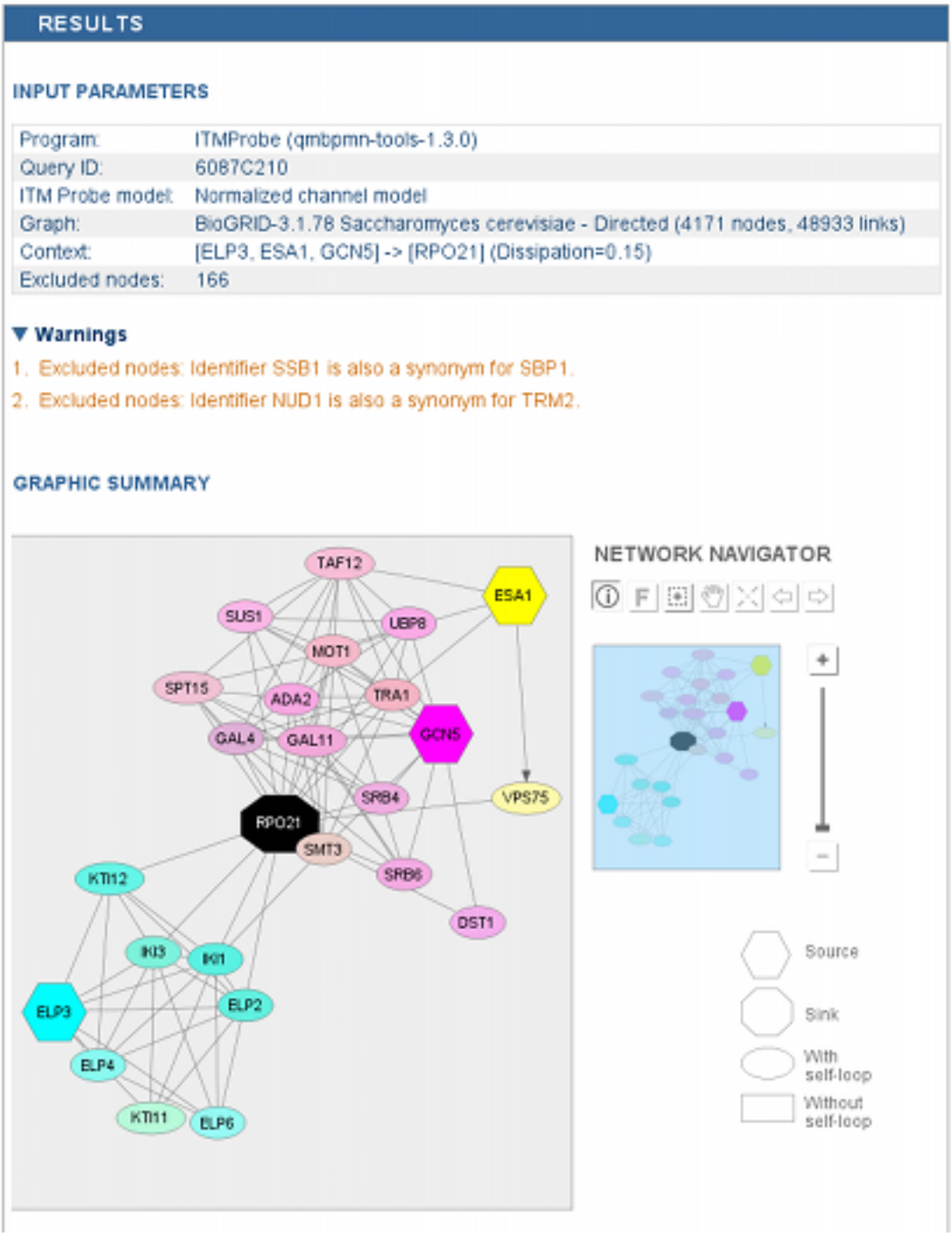
The **Color by** selection box specifies the criterion for coloring the nodes. When only a single source or sink are present only the *Total Visits* (for the channel and emitting models) and *Total Likelihood* (for the

absorbing model) criteria are available. When multiple boundary points are specified, it is possible to color the nodes according to the values associated with a single boundary point or according to the *Interference* at that node. Furthermore, if there are less than four boundary nodes, the *Intensities (color mixture)* option becomes available. In this case, each source (in the channel or emitting model) or sink (in the absorbing model) is assigned a basic CMY color and the coloring of each displayed node is a result of mixing the colors corresponding to its model values for each of the boundary points.

The **Scaling** selection box determines the scaling function applied to the node intensities before coloring, while the **Colormap** specifies the set of colors used to describe node values.

Each graphical layout can be rendered and saved in multiple formats (SVG, PNG, JPEG, EPS and PDF), selected in the **Image Format** box. The *SVG in Navigator* option produces an image of the ITM embedded into the *Network Navigator*, which allows scrolling and zooming the image.

2.3 Results

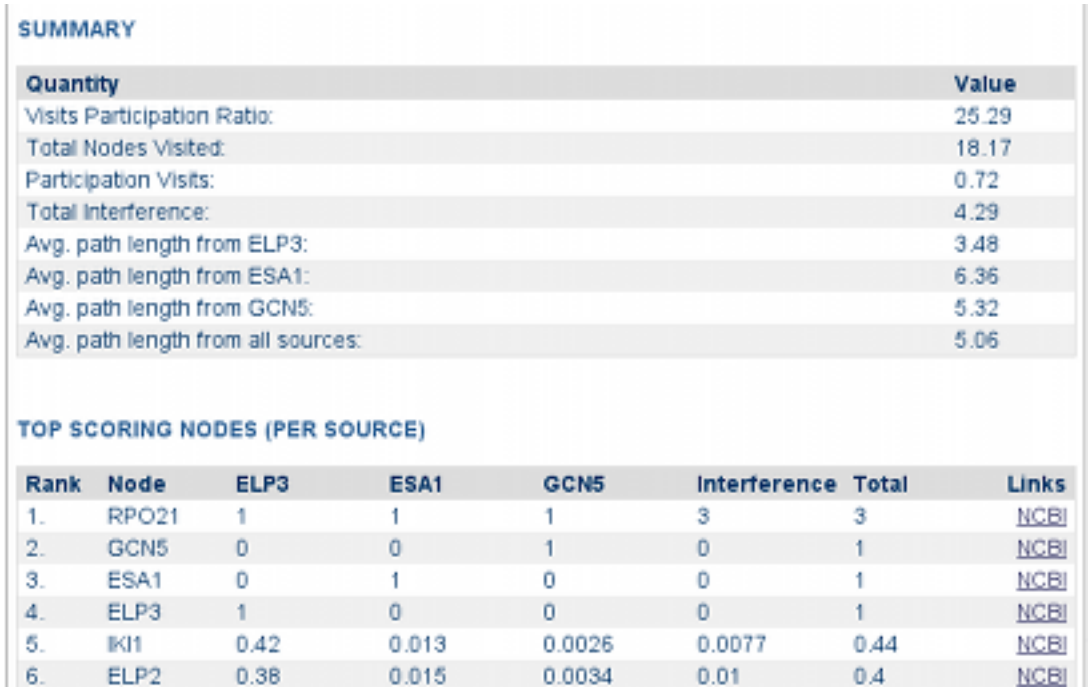


The top of the Results panel.

The Results panel consists of the query summary, ITM image, and the table of the ITM nodes shown in the image. Each row of the node table contains the protein name, the associated model values and the links to the NCBI Gene database containing the protein and gene details.

While the table shows only the nodes shown in the graphical display, the full model solution can be down-

loaded for further analysis in the CSV format by hitting the **CSV** button in the Display Parameters panel.



SUMMARY

Quantity	Value
Visits Participation Ratio:	25.29
Total Nodes Visited:	18.17
Participation Visits:	0.72
Total Interference:	4.29
Avg. path length from ELP3:	3.48
Avg. path length from ESA1:	6.36
Avg. path length from GCN5:	5.32
Avg. path length from all sources:	5.06

TOP SCORING NODES (PER SOURCE)

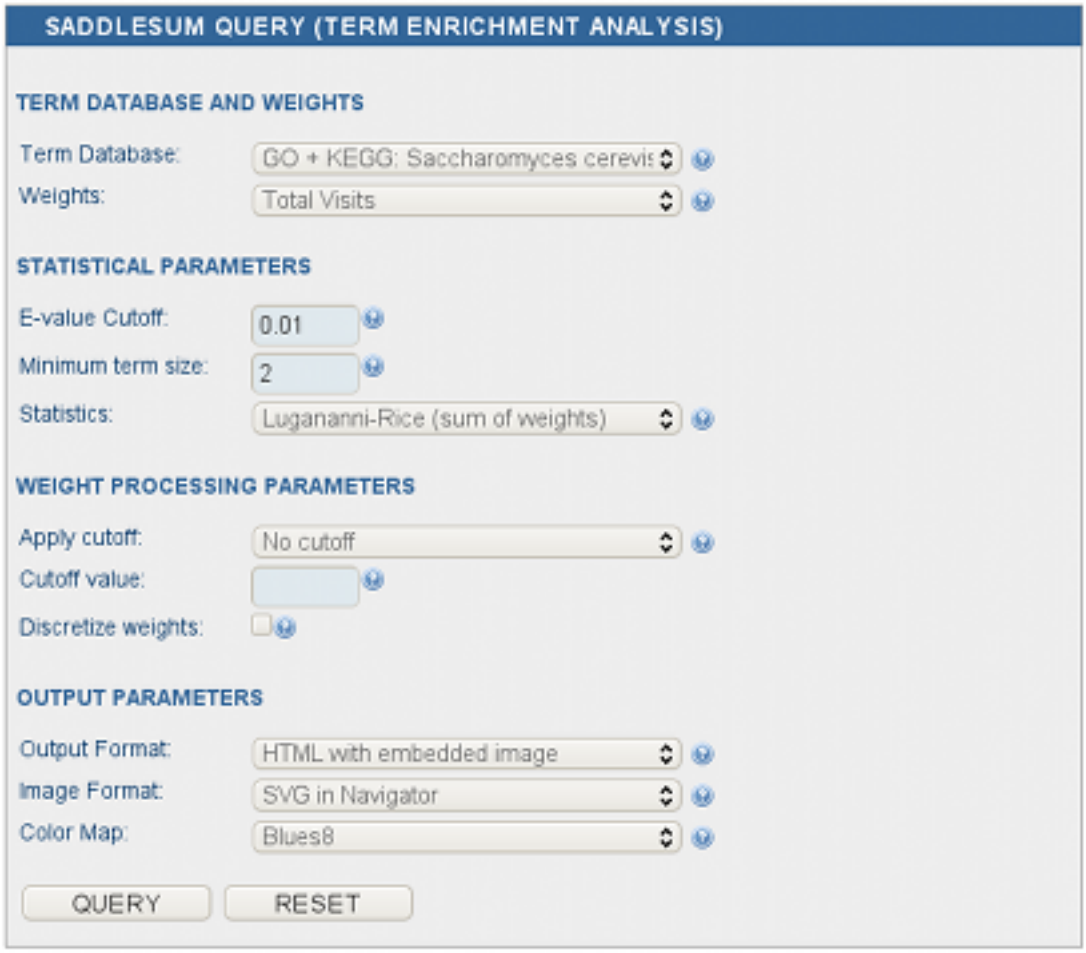
Rank	Node	ELP3	ESA1	GCN5	Interference	Total	Links
1.	RPO21	1	1	1	3	3	NCBI
2.	GCN5	0	0	1	0	1	NCBI
3.	ESA1	0	1	0	0	1	NCBI
4.	ELP3	1	0	0	0	1	NCBI
5.	IKI1	0.42	0.013	0.0026	0.0077	0.44	NCBI
6.	ELP2	0.38	0.015	0.0034	0.01	0.4	NCBI

The middle of the Results panel.

2.4 Term Enrichment Analysis

While it is possible to specify any proteins in the network as sources and sinks, not every context produces biologically meaningful results. In order to help the users to biologically interpret their results, we have developed SaddleSum, a term enrichment tool based on sum-of-weights scores, which gives accurate gene set enrichment P-values.

For a given set of weights over the network derived from an ITM, it finds the terms from SaddleSum's term databases (Gene Ontology and KEGG) that statistically best describe that set of weights. Every *ITM Probe* results page contains a query form allowing easy submission of the results to *SaddleSum*. For details about the query options, refer to the *SaddleSum* documentation.



SaddleSum query form

3 Cytoscape Plugin

CytoITMprobe is a Cytoscape plugin that provides interface to *ITM Probe* functionality. It works by querying *ITM Probe* either locally (using the standalone version from *qmbpmn-tools*) or remotely, through an HTTP request to a web server. The interface of *CytoITMprobe* is similar to that of the web version. However, *CytoITMprobe* significantly extends the features of the web version by supporting input of arbitrary networks and flexible manipulation and visualization of the query results. The entire results are stored as network and node attributes, allowing for their processing using other Cytoscape plugins with complementary functionality. This allows easy integration of *ITM Probe* into network-based data analysis workflows.

CytoITMprobe was developed for Cytoscape version 2.8. All the source code written a the NCBI is released into public domain.

Please watch this video demo, which complements the instructions here.

3.1 Downloading and Installing

It is possible to download *CytoITMprobe* either as a JAR file ready for installation as a Cytoscape plugin or as zipped source code archive. Both can be found on the NCBI FTP site (ftp://ftp.ncbi.nih.gov/pub/qmbpmn/CytoITMprobe/). Releases of *CytoITMprobe* share version numbers with *qmbpmn-tools*, starting with 1.4.

To install *CytoITMprobe*, copy the JAR file you have downloaded to the plugins subdirectory of your Cytoscape distribution or use the menu option *Plugins* → *Install Plugin from File* within Cytoscape. After successful installation was completed, you will see the **CytoITMprobe** entry in the **Plugins** menu.

3.2 Building from Source

The source code was mostly developed using NetBeans (http://www.netbeans.org/) and built using Apache Ant (http://ant.apache.org/). To build the JAR file, you first need to unzip the source distribution file, put `cytoscape.jar` onto your CLASSPATH and copy the file `build.xml.git` in the root of the distribution to `build.xml`. Then type:

```
ant package
```

to build the JAR named `CytoITMProbe.jar` in the root distribution directory.

3.3 Using CytoITMprobe

Starting plugin from Cytoscape

Start *CytoITMprobe* by choosing the Cytoscape menu entry *Plugins* → *CytoITMprobe* → *Query Form*. After reading the configuration file (see below), *CytoITMprobe* creates the query form and inserts it into the Cytoscape Control Panel (on the left of the Cytoscape window). You may need to resize the Control Panel to see the entire *ITM Probe* query form.

The query form provides a similar functionality to the query form of the *ITM Probe* web interface. Three action buttons can be found at its bottom:

- QUERY - to start running a *ITM Probe* query (see *Setting up a query* (page 16));
- RESET - to reset the form;
- CONFIG - to change *CytoITMprobe* configuration (see *Configuration* (page 15)).
- LOAD - to load *CytoITMprobe* results saved as attributes of the currently selected network (see *Saving and restoring results* (page 18)).

Configuration

Configuration of *CytoITMprobe* is handled though a separate dialog.

There are two possible configurations:

Web Query This option will configure *CytoITMprobe* to query the *ITM Probe* web service over HTTP. Choosing this method means that you are not required to download and install the *qmbpmn-tools* Python package and its dependencies. This is the recommended option.

Local Query This option will configure *CytoITMprobe* to run a local `itmprobe` program. Choosing this method means that you need to download and install the command-line version of *ITM Probe* and all its dependencies. Queries will be generally faster.

Configuring for web queries

Configuring *CytoITMprobe* to perform a web query requires clicking the "Web Query" radio button and entering the URL for the *ITM Probe* web services into the "Web Query URL" box.

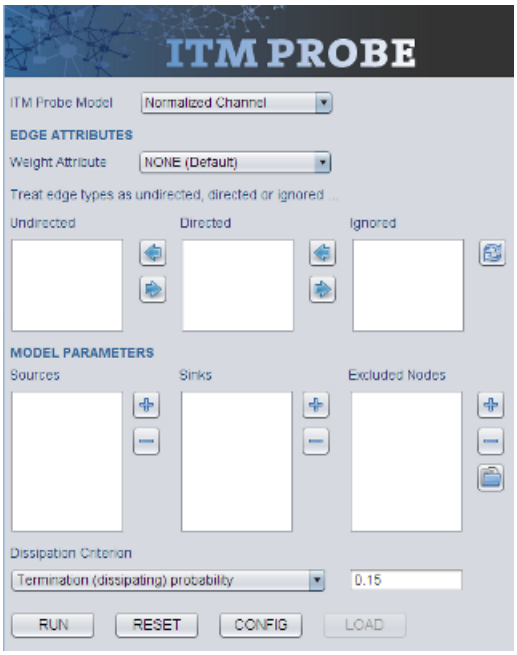
Note: When creating the initial configuration, *CytoITMprobe* will automatically set the URL of the default *ITM Probe* web service in the "Web Query URL" box and this need not to be changed unless the server address changes.

Configuring for local queries

Configuring *CytoITMprobe* to perform local queries requires significantly more work than configuring it to perform web queries. First, you need to download, and install the *qmbpmn-tools* package and all its dependencies (see *Python Package - qmbpmn-tools* (page 21)). In the configuration dialog you need to click the "Local Query" radio button and set the path for the `itmprobe` executable.

Setting up a query

To run an *ITM Probe* query, you need to specify via the query form a weighted directed graph, a model (absorbing, emitting or normalized channel) and some model parameters.



The graph connectivity is specified by selecting a Cytoscape network. In addition, you must assign a weight and a direction to each link. Edge weights can be set using the **Weight attribute** dropdown box, which lists all available floating-point edge attributes of the selected network and the default option (`NONE`). If the default option is selected, *CytoITMprobe* assumes the weight 2 for any self-pointing edge and 1 for all other edges. If an attribute is selected, the weight of an edge is set to the value of the selected attribute for that edge. Null attribute values are treated as zero weights.

Since Cytoscape edges are always internally treated as directed, you must also indicate the directedness of each edge type by placing it into one of the three boxes: `Undirected`, `Directed` or `Ignored`. Whenever a different Cytoscape network is selected, *CytoITMprobe* updates the query form and places all of the new network's edge types into the `Undirected`. You can then use arrow buttons to move some edge types to the `Directed` or `Ignored` category. Undirected edges are treated as bidirectional, with the same weight in both directions. Directed edges have a specified weight assigned only in the source-to-destination direction, with the opposite direction having the zero weight. Ignored edges have zero weight in both directions. Since Cytoscape allows multiple edges of different types between the same source and destination nodes, *CytoITMprobe* collapses multiple edges in each direction into a single edge by appropriately summing their weights.

The *ITM Probe* model is selected using the eponymous dropdown box. Depending on the selected model, you then need to select the sources and sinks. The absorbing model requires only sinks, the emitting model only sources, while the normalized channel model requires both. The **MODEL PARAMETERS** section contains three boxes for entering nodes: `Sources`, `Sinks` and `Excluded Nodes`. To add a node as a source or a sink, first select it in the current network and then press the + button next to the appropriate box. The node IDs for selected nodes will be added to the list in the box. To remove nodes from a list, select them and press the − button next to the box they are in.

Excluded nodes are specified by placing them into the `Excluded Nodes` box. There are two ways to do so: by selecting desired nodes in the network and using the + button and by loading a list of identifiers from a text file. The text file should contain identifiers delimited by one or more whitespace characters. For example, it can be formatted with one identifier per line. The identifiers should either be node IDs or `canonicalName` node attributes. Unrecognized attributes are ignored.

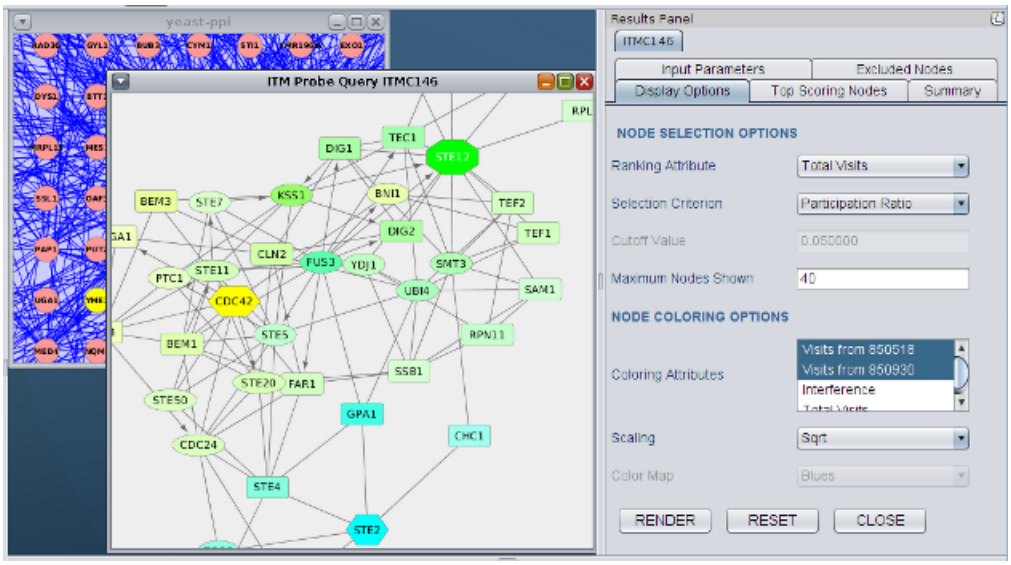
The choice of the model also determines the ways in which the dissipation coefficient can be set. You need to specify two items: the **Damping Criterion**, using a dropdown box, and the associated value. The default damping criterion is always `Termination (Dissipation) probability`, which requires a value between 0 and 1 and sets dissipation directly. In addition, the dropdown box will contain one or more model specific choices. The choices are described in detail in the section *Alternative Ways of Specifying Dissipation* (page 3).

To start a query, press the `QUERY` button. A progress dialog will appear and, after some time, the query results (or an error message) will appear.

Working with the results

Every completed *CytoITMprobe* query creates a viewer for its results embedded in Cytoscape Results Panel (located on the right of the Cytoscape main window when docked) and a new Cytoscape network that shows a subgraph of significant nodes (ITM subnetwork). To distinguish different queries, each query has a distinct name consisting of the prefix *ITM* followed by a letter denoting a model (*A* for absorbing, *E* for emitting, *C* for normalized channel), followed by a three-digit number (for example *ITME267*). The number is incremented every time a query is run up to 999 and then reset to 0.

The results viewer consists of five tabs. The tabs titled **Top Scoring Nodes**, **Summary**, **Input Parameters**, and **Excluded Nodes** contain information about the query and the results, while the **Display Options** tab contains a form that allows manipulation of the ITM subnetwork. The form controls two aspects of the subnetwork: composition (what nodes are selected and how many) and node coloring.



Note: The results viewer and the ITM subnetwork for each query query are linked. Destroying the ITM subnetwork will also remove the corresponding results viewer from the Results Panel.

ITM Subnetwork

Note: The visual style for the ITM subnetwork closely resembles the one presented by the web version. Of course, Cytoscape gives you the complete freedom to modify this look to your requirements.

Subnetwork nodes are selected through a **Ranking Attribute** dropdown box, which assigns a numerical value from *ITM Probe* results to each node. The nodes are ranked in decreasing order according to the ranking attribute and top scoring nodes are displayed as the ITM subnetwork. The number of top scoring nodes is determined by specifying a **Selection Criterion**, which can be simply a number of nodes to show, a cutoff value or the *Participation ratio*. Specifying a cutoff value x selects the nodes with their ranking attribute greater than x. The available choice for the ranking attribute depend on the *ITM Probe* model and the number of boundary points. For the emitting and normalized channel model, you can select visits to a node from each source, the sum of visits from all sources or *Interference*. For the absorbing model, the available attributes are absorbing probabilities to each sink and the total probability of termination at a sink. The values for all attributes for the subnetwork nodes are displayed in the **Top Scoring Nodes** tab.

The colors of the subnetwork nodes are determined by selecting one or more **Coloring attributes**, a **Scaling Function** and a **Color Map**. The list of coloring attributes is the same as the list of ranking attributes but you can select up to three coloring attributes. If a single attribute is selected, node colors are determined by the selected eight-category *ColorBrewer* color map. Otherwise, they are resolved by color mixing: each coloring attribute is assigned a single basic color (cyan, magenta or yellow), and the final node color is obtained by mixing the three basic colors in proportion to the values of their associated attributes at that node. The scaling function serves to scale and discretize the coloring attributes to the ranges appropriate for color maps.

Note: *CytoITMprobe* copies uses the nodes from the original network in the newly created ITM subnetwork. Thus, when selecting a node in ITM network, you are able to see all of it attributes in the original network. However, the edges between nodes are created from scratch. This is because Cytoscape allows more than one edge between two nodes, although they must be of different type. On the other hand, *CytoITMprobe* appropriately sums the weights of all the different edges between two nodes and thus cannot distinguish between them. Furthermore, the ITM subnetwork appears less crowded with only a single edge between a pair of nodes.

Saving and restoring results

Each *CytoITMprobe* query stores its results by setting node and network attributes in the network used for the query. This network and its attributes can be saved through Cytoscape, and then be reloaded in a different session. To restore the corresponding results panel, select a network containing the results and hit the `LOAD` button on the query form.

Note: The `LOAD` button will be enabled only if the selected network contains saved results and these results are not already shown in Results Panel.

Alternatively, the ITM Probe results stored in a network can be exported to tab-delimited text files through the Cytoscape **Export** menu: *File* → *Export* → *ITM Probe Results as TAB File...*. The tab-delimited export

file contains all information necessary to restore the results except the original network and is easy to read by both humans and custom programs.

The results from exported tab-delimited files can be restored to a network through Cytoscape **Import** menu. You need to select a network and select a tab-delimited file using the *File* → *Import* → *Import ITM Probe Results from TAB File...* menu item. CytoITMprobe will attempt to load the results as network attributes and, if successfull, display the corresponding subnetwork and viewer. The imported ITM will be assigned a new label, as if it originated from a direct *ITM Probe* query.

Note: Only the results for the nodes in the selected network whose IDs match the IDs from the imported file will be loaded. Thus, it is possible to transfer an ITM from one network to another by exporting it and then importing it, as long as the two graphs have nodes with IDs in common.

Additional Node Attributes

Since the *ITM Probe* query results are saved as Cytoscape attributes of the original network, you can arbitrarily modify them through Cytoscape. The changes made in this way are reflected in the results viewer and ITM subnetwork after you press the `RESET` button in **Display Options** tab.

> **Warning:** Editing raw results in this way can cause unpredictable effects to the functioning of *CytoITMprobe*. Make sure you have the original results saved before any editing so you can restore them if problems occur.

Using the *CytoITMprobe* attribute nomenclature, you can create additional attributes to be used for ranking or coloring. Consider the following usage example. You have ran an emitting model query with three sources, `S1`, `S2`, and `S3`, and obtained the results in a viewer labeled `ITME243`. At the end of the run, *CytoITMprobe* created the attributes `ITME243[S1]`, `ITME243[S2]` and `ITME243[S3]` for the nodes of the input network and saved the results as their values. Using Node Attribute Browser within Cytoscape, you can create a new floating-point node attribute with a label `ITME243[avgS1S2]` and fill it with an average of `ITME243[S1]` and `ITME243[S2]` using Function Builder.

After resetting the Display Options form, an item `Custom [avgS1S2]` will be made available for selection as a ranking or coloring attribute. This gains you the ability to reinterpret `S1` and `S2` as if they were a single source of equal weight as `S3`. You can use the same procedure to combine the results of queries with different boundaries and display them together on the same subnetwork.

Example

Here is a step-by-step example to help you get started with *CytoITMprobe*. It involves the yeast protein-protein interaction network and the pheromone signal transduction pathway. The same example can be run by clicking the `EXAMPLE` button on the web form.

Note: This tutorial assumes basic familiarity with Cytoscape and that the *CytoITMprobe* plugin is already properly installed.

---

1. Download a zipped archive `CytoITMprobe-example.zip`, containing a yeast protein-protein interaction network in SIF format, a list of excluded nodes and a Cytoscape node attributes file from ftp://ftp.ncbi.nih.gov/pub/qmbpmn/CytoITMprobe/example/.
2. Create a new directory and extract the archive in it.
3. Start Cytoscape.
4. Activate *CytoITMprobe* from plugins menu. You may need to enlarge the Control Panel to see the entire query form.
5. Using the Cytoscape import dialog accessible from *File* → *Import* → *Network (All file types)...*, import the yeast PPI network from the extracted `yeast-ppi.sif` file.
6. Select the new network and import the `canonicalName` attribute from the extracted `canonicalName.NA`. You can use VizMapper to use the `canonicalName` attribute as node label rather than ID.
7. Make sure that the imported network is selected.
8. Activate the *CytoITMprobe* query form. Select `Normalized Chanel` as the *ITM Probe* model. Move edges of type `d` into `Directed` box using the  buttom next to the `Undirected` box. Leave the dissipation criterion at the default value: `Termination (Dissipation) probability` with the probability 0.15.
9. Click on the `Configure search options` toolbar button and select `canonicalName` in the **Select attribute** dropdown box. This will allow you to search for network nodes using their cannonical names rather than by their numerical IDs.
10. Use the Cytoscape search facility to add the nodes `STE2` and `CDC42` as sources and `STE12` as a sink. In each case, start by typing the desired node name in the search box; there will be a single hit, which will become selected. Use the  button next to the `Sources` or `Sinks` box to add the selected node to the appropriate list.
11. Press the  button next to the `Excluded Nodes` box and load the suggested excluded nodes from the file `yeast-excluded.txt`, which was extracted earlier.
12. Hit the `QUERY` button. After a while you will see a new network created and a new tab in the Results Panel.
13. You can browse the result tables on the right or try different visualizations of the results using the **Display Options** form. By default, the ITM subnetwork shows at most 40 nodes using `Total Visits` as the selection criterion. The default coloring attribute is also `Total Visits`. You can change to a mixed color view by selecting more than one (but no more than three) attributes in the **Coloring Attributes** list.
14. This finishes the step-by-step tutorial. You can now explore the rest of the *CytoITMprobe* interface on your own and perhaps rerun the query with different parameters.

4 Python Package - qmbpmn-tools

qmbpmn-tools is a collection of Python libraries and scripts for network analysis developed by the QMBP group at the NCBI. It consists of the *ITM Probe* library and executable script, the routines for managing *ITM Probe* and *SaddleSum* web sites, a package manager and miscellaneous libraries and utilities.

Warning: *qmbpmn-tools* is in heavy development and may change in many ways. We do not support this package but only release it as a reference implementation. This documentation is therefore incomplete and should be considered as short installation notes rather than as user's manual. The best way to obtain detailed information is to actually read the source code.

4.1 Prerequisites

qmbpmn-tools requires a number of standalone programs and Python packages. All are open source and can be downloaded and installed on UNIX, Windows and Mac OSX machines. Please consult the documentation of individual packages for information of how to install them. The prerequisites are:

- Python (http://www.python.org/) 2.6 or 2.7 complied with SQLite (http://www.sqlite.org/) support. Python 3 is not yet supported.
- SaddleSum standalone program for *SaddleSum* website functionality.
- Graphviz (http://www.graphviz.org/) graph visualization software. **Not necessary for CytoITMprobe**.
- Numpy and Scipy (http://www.scipy.org/) libraries of scientific tools for Python. Numpy should be version 1.3 or higher, while Scipy should be version 0.7 or higher.
- SciKits.umfpack (http://scikits.appspot.com/umfpack) Python package (**Optional but highly recommended**). It requires UMFPACK (http://www.cise.ufl.edu/research/sparse/umfpack) library for solving sparse systems of linear equations. Ensure that UMFPACK is compiled with good BLAS, otherwise ITM Probe performance will greatly suffer.
- Jinja2 (http://jinja.pocoo.org/2/) template engine. **Not necessary for CytoITMprobe**.
- Sphinx (http://sphinx.pocoo.org/index.html) Python documentation generator (version 1 or higher). Sphinx additionally requires docutils (http://docutils.sourceforge.net/) and Pygments (http://pygments.org/) packages. **Not necessary for CytoITMprobe**.

4.2 Downloading and Installing

The source code is available on the NCBI FTP site (ftp://ftp.ncbi.nih.gov/pub/qmbpmn/qmbpmn-tools/) as a tar.gz archive. Extract the archive to a temporary directory, and install it in the standard way:

```
python setup.py install
```

The setup program will install the package and also place several executable scripts in your path.

4.3 Components

qmbpmn-tools source package consists of four subdirectories:

- **scripts/** contains executable programs (i.e. Python scripts) such as
  - `itmprobe` - standalone *ITM Probe*,
  - `qmbpmn-deploy` - a script to deploy the website,
  - `qmbpmn-server` - a rudimentary deployment server,
  - `qmbpmn-datasets` - package manager,
  - `cluster` - a script to cluster vectors using the method described in Hamaneh MB, Yu YK. PLoS One. 2014 Oct 31;9(10):e110936

Most scripts have some help available. Use:

```
<script-name> -h
```

to print help to the screen.

- **ITMProbe/** contains *ITM Probe* core library
- **web/** contains Python and resource files for the *ITM Probe*, *SaddleSum* and *DeCoaD* web sites.
- **common/** contains shared subpackages as well as code obtained from other open-source projects.

5 Additional Information

5.1 License

All code for *ITM Probe* (*qmbpmn-tools*) and *CytoITMProbe* written at the NCBI is released into Public Domain. The licenses of external components are indicated in the source packages.

5.2 References

Further details about *ITM Probe* can be found in the following references:

1. A. Stojmirovic and Y.-K. Yu. Information flow in interaction networks. *J. Comput. Biol.*, **14** (8):1115-1143, 2007.
2. A. Stojmirovic and Y.-K. Yu. ITM Probe: analyzing information flow in protein networks. *Bioinformatics*, **25** (18):2447-2449, 2009.

5.3 Credits

- Aleksandar Stojmirovic and Yi-Kuo Yu designed the study, conducted the research and wrote the papers;
- Aleksandar Stojmirovic wrote the code for the *qmbpmn-tools*.
- Alexander Bliskovsky and Aleksandar Stojmirovic wrote together the code for *CytoITMProbe*.

5.4 Acknowledgments

ITM Probe (*qmbpmn-tools* package) is written in the Python (http://www.python.org/) programming language and relies on several open-source components:

- Numpy and Scipy (http://www.scipy.org/) libraries of scientific tools for Python;
- UMFPACK (http://www.cise.ufl.edu/research/sparse/umfpack) library for solving sparse systems of linear equations;
- Graphviz (http://www.graphviz.org/) graph vizualization software;
- Sphinx (http://sphinx.pocoo.org/index.html) Python documentation generator;
- Jinja2 (http://jinja.pocoo.org/2/) template engine.

The web browser client code uses various Javascript routines from the NCBI. ECMAScrpit code and widgets for the SVG 'Network Navigator' were taken from Carto:Net (http://www.carto.net/).

All discrete network image color schemes were taken from the www.ColorBrewer.org (http://www.ColorBrewer.org) site by Cynthia A. Brewer, Geography, Pennsylvania State University.

Protein interaction graphs for the web version were constructed using data from the BioGRID (http://www.thebiogrid.org/) database.

CytoITMProbe uses Apache HttpComponents (http://hc.apache.org/) library for HTTP requests and Google gson (http://code.google.com/p/google-gson/) for manipulating JSON files. The icons used in the graphical interface were taken from the Tango Icon Library (http://tango.freedesktop.org/Tango_Icon_Library), version 0.8.90.

We are grateful to Zvezdana Stojmirovic for help with the graphic design the *ITM Probe* web pages and interfaces.

5.5 News and Updates

- **01-Sep-2017.** *qmbpmn-tools* 1.5.4 released - a bug fix and some other small changes.
- **03-Jul-2013.** *qmbpmn-tools* 1.5.3 released - minor changes and bug fixes.
- **12-Mar-2012.** *CytoITMprobe* 1.5 released.
- **02-Mar-2012.** *qmbpmn-tools* 1.5.2 released - a bug fix for *ITM Probe*.
- **07-Feb-2012.** *qmbpmn-tools* 1.5.1 released - minor changes and bug fixes for *ITM Probe*.
- **14-Dec-2011.** *qmbpmn-tools* 1.5.0 released - minor changes mostly related to *SaddleSum*.
- **11-Jul-2011.** *ITM Probe* 1.3.0 released, with updated web forms. It is now possible to retrieve an old job using its query ID.

6 Glossary

Absorbing model Corresponds to absorbing Markov chains. Contains only sinks as boundary and for each node in the network evaluates the likelihood for a random walk starting at that node to terminate at each sink while avoiding all other sinks (hence the sinks can be thought of as competing for the flow). The total likelihood at each node is the sum of the individual likelihoods for all sinks. Due to dissipation, the total likelihood at each node can be much less than 1, especially for the nodes far from any of the sinks.

Channel model Contains both sources and sinks and combines emitting and absorbing models. It reports the total expected number of visits to any node in the network from a random walk originating at each source and terminating at any of the sinks (i.e. not dissipating).

Context A set of parameters that defines the environment for a random walk in an interaction graph. It includes boundary nodes (sources and/or sinks), rate of dissipation and the set of excluded nodes. *ITM Probe* evaluates the context to describe the information flow it represents.

Dissipation rate Also known as termination/stopping probability. A parameter between 0 and 1 describing the proportion of random walks that leave the network (evaporate) at each step. The higher the dissipation rate, the more likely it is for the random walk to terminate close to its origin and thus have only local effects. In the context of the channel model, the dissipation rate controls the likelihood for the random walk to visit the nodes away from the shortest paths from sources to sinks – dissipation close to 1 means that only the nodes on shortest paths will be visited.

Emitting model Contains sources on the boundary. For each node in the network evaluates the average number of time the node is visited by a random walk originating at each source. Random walks terminate when reaching any source or dissipating.

Excluded node A protein node that is excluded from consideration during a run of *ITM Probe*.

In some cases, proteins with a large number of non-specific interaction partners might overtake the true signaling proteins in the information flow modeling. Therefore, *ITM Probe* allows users to specify nodes to exclude from the network. For the yeast network the nodes excluded by default include cytoskeleton proteins, histones and chaperones, since they may provide undesirable shortcuts.

Note that excluded nodes are treated as terminating points for random walks: the edges leading into them are not deleted but any random walk entering and excluded node evaporates instead.

Interaction network A (weighted, directed) graph whose nodes (vertexes) represent agents and edges (links) are interactions. Thus, two agents are linked if they interact in some way. The weight of a link corresponds to the strength of the corresponding interaction.

Interference A concept used in the context of the emitting and the channel model to describe the measure of overlap between visits of each node by random walks originating at different sources. Large interference implies large overlap between flows from different sources while small interference means little overlap.

ITM **I**nformation **T**ransduction **M**odule. The set of most significant nodes (with respect to the number of visits, in the case of the emitting and channel models, or the absorption probability, in the case of the absorbing model) resulting from the query context of *ITM Probe*.

Participation ratio Used for the emitting and the channel model. Gives approximate number of nodes that have a significant (that is, much larger than ordinary) total number of visits.

Protein interaction network An interaction network where nodes correspond to cellular proteins. The edges may represent a variety of interactions, for example: physical (protein A binds protein B), metabolic (A and B catalyze reactions involving the same chemical), genetic (A and B are expressed together) or biochemical (A post-translationally modifies B). It is also possible to include more than one type of interaction in the network, depending on the problem being modelled.

Random walk A mathematical concept involving an entity that moves about a given space in a random fashion. In the context of graphs, a random walk describe a process where a 'walker' moves from one vertex into another with a probability proportional to the weight of the edge connecting them. This process is equivalent to a Markov chain on the vertex set. A random walk starts at a node in a network and moves about visiting different nodes until it terminates. It can terminate either at a boundary point (source or sink) or by leaving the network due to dissipation.

Sink A destination of a random walk.

Source A point of origin of a random walk.

Termination probability see *Dissipation rate*.

Total likelihood see *Absorbing model*.

Total visits Used in the context of the emitting and the channel model to denote the total number of visits from all sources.